ARI Research Note 97-05

The Utility of the Training and Evaluation Outline Data Base as a Performance Measurement System at the Joint Readiness Training Center

Gene W. Fober
U.S. Army Research Institute

Infantry Forces Research Unit
Scott E. Graham, Chief

March 1997

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

Technical review by

Robert J. Pleban
William R. Rieger

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED (from. . . to)
1997, March	Final	October 1995-September 1996

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
The Utility of the Training and Evaluation Outline Data Base as a Performance Measurement System at the Joint Readiness Training Center	5b. PROGRAM ELEMENT NUMBER 0603007A
6. AUTHOR(S) Gene W. Fober (ARI)	5c. PROJECT NUMBER A793
	5d. TASK NUMBER 2127
	5e. WORK UNIT NUMBER H01
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: PERI-IJ P.O. Box 52086 Fort Benning, GA 31995-2086	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 5001 Eisenhower Avenue Alexandria, VA 22333-5600	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Research Note 97-05

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT *(Maximum 200 words)*:

The purpose of this research was to examine the Training and Evaluation Outline (T&EO) data base for utility as a performance measurement system. Previous research had determined that the data base was of limited value for making empirical analyses of the Joint Readiness Training Center (JRTC) performance data. Based on recommendations from the previous findings, JRTC changed the performance measurement system. The changes included the introduction of a five-point rating scale and a reduction in the number of rated items. The current research was conducted to determine whether these additions increased the utility of the T&EO data base as a feedback and performance measurement system. T&EO data were analyzed at battalion task force, company, and platoon levels for nine rotations at the JRTC. It was found that the T&EO data base still lacks the reliability required to provide useful feedback to units or to provide researchers with useful information on unit trends. Although tasks differed statistically, the usefulness from a practical standpoint is limited because the range of scores is too narrow. Potential users of the data base are cautioned not to make conclusions based solely on statistical significance. Recommendations to improve the data include reducing the rating categories to more general levels and placing a greater emphasis on the importance of a quality performance measurement system. A method to reduce the number of rating categories using subject mater experts was introduced as one way to improve the performance measurement system.

15. SUBJECT TERMS

Performance measurement Training and Evaluation Outline (T&EO)

Joint Readiness Training Center (JRTC)

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	24	Carol J. Bryan DSN835-5589

THE UTILITY OF THE TRAINING AND EVALUATION OUTLINE DATA BASE AS A PERFORMANCE MEASUREMENT SYSTEM AT THE JOINT READINESS TRAINING CENTER

CONTENTS

LIST OF TABLES

Page

LIST OF TABLES (Continued)

LIST OF FIGURES

THE UTILITY OF THE TRAINING AND EVALUATION OUTLINE DATA BASE AS A PERFORMANCE MEASUREMENT SYSTEM AT THE JOINT READINESS TRAINING CENTER

Background

Combat Training Centers (CTCs) have been a vital training link in ensuring unit combat readiness. Units deploy to CTCs to conduct realistic missions against dedicated opposing forces. The performance feedback and training recommendations resulting from a CTC rotation aid the unit in the assessment of its combat readiness. Performance feedback and training recommendations come from a variety of sources. Most information can be traced back to the Observer/Controllers (O/Cs) assigned at various elements and echelons to provide the critical assessment of performance resulting in unit performance feedback.

Besides providing units with performance feedback and training recommendations, data from the CTCs may be used by researchers to investigate various training issues. Although individual units benefit from the immediate performance feedback, the army as a whole benefits from data collected during many rotations. Rotational data gathered over time may provide insights into an overall army training strategy based on consistent patterns of performance. Fober, Dyer, & Salter (1994) documented several types of available CTC data and possible research methods associated with the data.

One available data source which is unique to the Joint Readiness Training Center (JRTC) is the training and evaluation outlines (T&EOs). T&EOs consist of Infantry tasks, subtasks, and subtask standards found in the Infantry Mission Training Plans (MTP) (e.g., DA, ARTEP 7-8-MTP, 1988). Tasks are rated as T (trained), P (needs improvement), or U (untrained). However, task standards, subtasks, and subtask standards are rated on only a GO/NOGO scale. The MTP tasks were transferred to a JRTC document called the "Greenbook". O/Cs were provided with a Greenbook for each mission. All possible tasks based on the mission were contained within the Greenbook.

The T&EOs were developed to be training and evaluation aids. They are used extensively by units training at home station. T&EOs are used to plan training and to determine progress during training exercises. They are ideal for commanders because they allow commanders to determine future unit training requirements. This is accomplished by using the T&EO checklists during training and using the evaluation results as feedback to the individual unit leaders.

Previous Research Findings

As a performance measurement system at the JRTC, T&EOs proved to be too detailed and cumbersome to provide useful unit feedback (Fober, 1993). O/Cs have more responsibilities than just filling out the T&EOs. For example, O/Cs must conduct an after action review (AAR) with the observed element on completion of each mission. The format of the T&EOs is such that

filling them out does not assist O/Cs in the AAR process. In addition, tasks performed under simulated combat conditions are not always performed in sequence. For example, many of the subtasks may not apply to the specific combat conditions, whereas at home station tasks can be isolated and performed sequentially. Another related problem is that competing demands on O/C's time resulted in checklist completion long after the tasks had been performed. In spite of the problems with using T&EOs as a performance measurement system at JRTC, the data were collected for about five years. The data were coded by rotation, mission, element, etc. (see Fober, 1993; Nichols, 1991).

Fober (1993) examined the data to determine if the T&EO data base was useful as a performance measurement system, both from a unit feedback and a research perspective. The research examined company and platoon results from five consecutive rotations. Based on the results of that study, Fober (1993) recommended changes to the JRTC T&EO data collection and coding procedures. These changes were recommended only for the JRTC T&EO data base and not for home station training. T&EOs contained in the MTPs serve their intended purpose during home station training because they are a means for units to assess future training needs. However, T&EOs have not been adequate as a performance measure because the majority of the tasks were rated "Untrained" with subtasks and subtask standards being rated "NOGO". Thus, the additional detail of the subtasks and the subtask standards did not provide any specific direction for future training other than train everything.

As a result of Fober's (1993) concerns, two major changes to the T&EO data base were adopted: an expanded rating scale (see Table 1) and a reduction in the number of items rated. These two changes addressed some of Fober, Dyer, and Salter's conerns with the JRTC T&EO measurement system. It was thought that expanding the rating scale would provide a means for researchers to determine differences in task performance. Part of the lack of performance differences found among the tasks may have been due to the two- and three-point scales (Fober, 1993). The reduction in the number of items rated was suggested as a way to reduce the overall O/C workload. The hypothesis was that excessive O/C workload may have contributed to the lack of discrimination among tasks and their sub elements. This assumption was based on examination of the data and interviews with O/Cs who indicated that the Greenbooks were not high on their list of post rotation priorities (Fober, 1993). It was expected that rating only at the task and subtask levels would reduce the overall O/C workload. The loss of detail associated with fewer ratings could be offset by the introduction of the five-point rating scale.

Although there were changes in the T&EO measurement system, the methods of completing the ratings remained similar to what was reported in Fober (1993). O/Cs were issued pocket-sized books, also known as Greenbooks, containing possible tasks, by mission, for the element they were observing. O/Cs rated the tasks and subtasks using the five-point scale. The words for the scale adopted were based on reviews of take home packages and interviews with O/Cs (e.g., Unit demonstrated technical and tactical proficiency required to perform task to standard). The task standards and subtask standards were not rated, but were still provided within the greenbooks as reference only. Again, each task was only rated at the overall task level and at the subtask level. Because some subtasks have as many as 20 subtask standards, rating at

only the subtask level greatly reduced the number of rated items. An example of the subtasks accociated with a platoon task ("Prepare for Combat") can be found at Table 2.

Table 1

New T&EO Rating Scale Contained in JRTC Greenbooks

Rating Scale	Verbal Description
1 - Poor	Unit completely lacked technical and tactical proficiency to perform this task to standard.
2 - Weak	Unit attempted to perform task but lacked technical and tactical proficiency to meet all standards.
3 - Adequate	Unit demonstrated technical and tactical proficiency required to perform task to standard.
4 - Good	Unit demonstrated technical and tactical proficiency to perform task and exceeded some standards.
5 - Excellent	Unit demonstrated technical and tactical proficiency to perform task and exceeded most standards.

A preliminary analysis of the updated T&EO data base indicated some promise for researchers desiring to use the T&EO data base as a criterion measurement system (Fober, Dyer, & Salter, 1994). Tentative conclusions were made based on the data from three rotations. Based on this sample, it appeared that the O/Cs were using the entire five-point scale, thereby increasing the possibility for determining task performance trends. The limited amount of data did not allow an examination of additional issues and concerns identified by Fober (1993). It was recommended that a follow-up study was needed using a greater sample size.

Purpose of the Present Study

The present study examines nine recent rotations which employed the new rating scale. Issues and concerns from previous research (Fober, 1993; Fober, Dyer, & Salter, 1994) were re-examined to determine whether the changes made to the JRTC T&EO rating system increased the utility of the data for research and unit feedback purposes. One of the changes to the rating system was the reduction in the number of items to rate (i.e., O/Cs rate only tasks and subtasks). This reduction in O/C workload may impact on the tendency of some O/Cs to rate everything the same (i.e., O/C biases or tendencies, as reflected in halo or floor effects). Another change, the introduction of the five-point scale might increase the possibilities for discrimination between tasks. That is, if performance differences between tasks exist, then an increased scale is more likely to reveal those differences than the old "GO/NOGO scale.

Table 2

Subtasks Associated with Platoon "Task Prepare for Combat"

Task 611: Prepare for Combat
1. The platoon leader receives the mission (Step 1 - TLP (Troop Leading Procedures) from the company commander.
2. The platoon leader performs mission analysis.
3. The platoon leader completes mission analysis.
4. Platoon leader issues a warning order to platoon sergeant and squad leaders (Step 2, TLP).
5. Platoon performs readiness, maintenance, and functional checks under leader supervision.
6. Vehicles are combat loaded in accordance with (IAW) the standard operating procedure (SOP) or warning order.
7. All personnel test-fire weapons, if the situation permits.
8. Platoon leader makes a tentative plan (Step 3, TLP).
9. Platoon initiates movement (Step 4, TLP) as required for quartering party, selected elements, or the entire platoon (T&EO 7-3/4 - 1025, Move Tactically).
10. The platoon conducts recon (Step 5, TLP).
11. Platoon leader completes plan (Step 6, TLP) based on METT-T (Mission, Enemy, Troops available, Terrain, Time), intel from recon, and commander's guidance.
12. Platoon leader issues operations order (OPORD) to subordinate leaders (Step 7, TLP).
13. Platoon leader conducts coordination.
14. The platoon leader supervises mission preparation.
15. The platoon plans sustainment of combat operations.
16. Platoon performs continuous recon during the operation.
17. Platoon monitors actions of higher, adjacent, and supporting units.
18. Platoon leader issues orders or modifies original plan.
19. Platoon leader issues fragmentary orders (FRAGOs) to the platoon and attached elements.
20. The platoon reacts to orders from higher headquarters.
21. The platoon coordinates actions with friendly units during the operation.
22. Platoon Headquarters reports combat critical information to higher, adjacent, and supporting units.
23. Platoon Headquarters disseminates information to the platoon.

Note. The new performance measurement system required O/Cs to rate only the overall task and the above subtasks on the five-point scale. See DA, ARTEP 7-8-MTP, 1988 for the entire task including subtask standards as well as other platoon tasks.

Examination of T&EO Data Base

Scale Discrimination

One issue concerning the rating scale is whether or not the O/Cs use the scale as it is intended. In the preliminary analysis which summarized data over three rotations (Fober, Dyer, & Salter, 1994), it appeared that O/Cs were using the full scale. There was a tendency for more ratings at the middle and lower end of the scale, but large numbers of ratings over the entire scale was cause to be optimistic. The authors did point out that the limited amount of data made it difficult to determine whether individual O/Cs were using the full scale. A summary of the whole data base might average out any individual tendancies. Therefore, a detailed examination of the data can reveal to what extent individuals may restrict their use of the rating scale. If it is found that a large number of O/Cs are restricting their use of the rating scale, then there may be concerns about the utility of the data as a performance measurement system.

This section examines the issue of how O/Cs use the rating scale. First, overall subtask ratings from nine rotations were summarized to determine whether the entire scale was used. Company and platoon subtask ratings were also summarized from the same nine rotations to determine whether O/Cs at different echelons were using the entire scale. Finally, the data were examined by element over the nine rotations to determine possible trends related to how individual O/Cs used the rating scale.

Overall subtask ratings. Figure 1 presents a summary of all the subtask ratings over all echelons for the entire nine rotations. No statistical analyses were conducted on the overall subtask ratings. A visual examination of Figure 1 reveals that the entire scale contained ratings. There appear to be sufficient data within each of the five rating categories to allow researchers to determine trends in task performance. One possible concern is that more than half of the ratings (n=84,866) were in the "adequate" category. The least used rating category was "excellent" which contained 5,868 ratings. Thus, the results presented in Figure 1 are cause for cautious optimism. Given the large amount of data, researchers might be able to perform statistical analyses which would assist in determining overall training trends. However, another possible concern is that summarizing over all echelons created the appearance that the scale was being used by the majority of O/Cs for its intended purpose. Therefore, the next step was to obtain summaries of battalion, company and platoon subtask ratings to determine possible differences in the way O/Cs from different echelons rate subtasks.

Subtask Ratings by Echelon. The summaries of the subtask ratings for battalion task force, company and platoon are presented in Figure 2. Statistical comparisons were not made because the large number of cases would almost certainly result in statistical significance. From a practical standpoint a visual examination of the pattern of results is more beneficial. Examination of ratings over the three echelons reveal similar patterns to those found in the overall subtask ratings. Again, the results are encouraging because the range of ratings encompass the entire scale. More detailed analyses are required to determine whether individual

O/Cs are using the entire scale and whether discrimination of individual task performance can be discerned from the present data.

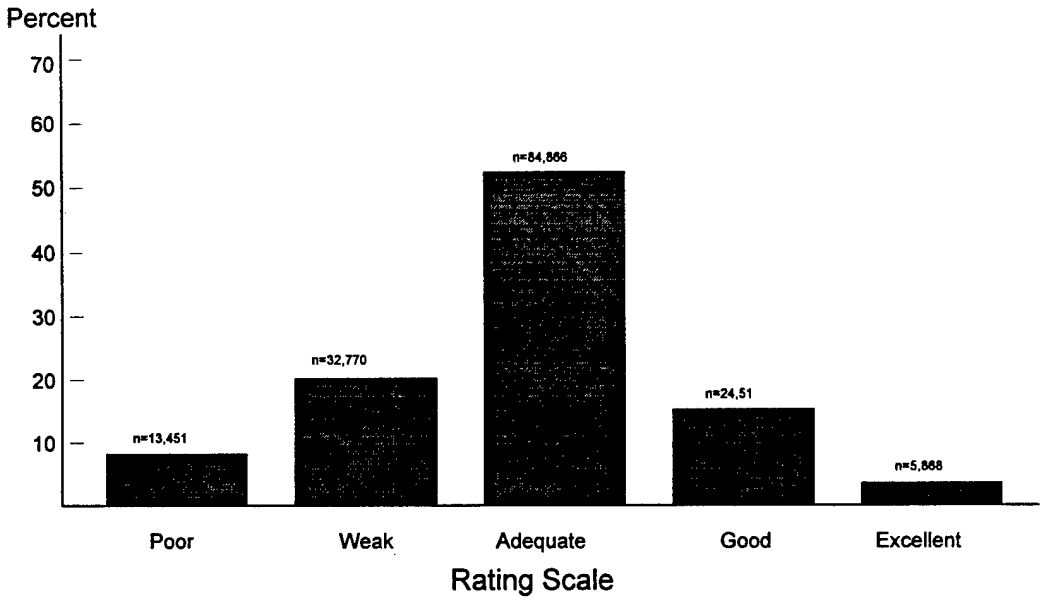

Figure 1. Percent subtask ratings for nine rotations (overall).

Task Discrimination. One of the main problems with the original rating system from a performance measurement standpoint was the scale (i.e., GO/NOGO). This resulted in little task discrimination because the majority of subtask ratings were "NOGO". Thus, one intended purpose of the data base could not be performed. A useful performance measurement system must enable the researcher to discriminate among tasks. The measurement system must be of sufficient fidelity and sensitivity to track a task over time to determine the impact of new training, new equipment, or doctrinal changes. It must also enable researchers to determine which tasks typical units need to improve upon. If researchers can not perform these basic functions, then the performance measurement system does not serve its intended purpose.

To determine whether O/Cs rated tasks differently, the most frequently performed tasks were selected from both platoon and company data. Comparisons were made based on overall task scores assigned by the O/Cs. Means were calculated by task for each rotation (see Table 3 for platoon task means). For example, a task mean for a platoon task in Rotation 1 included all iterations of the task and all platoons which were rated on the task. Then a Task by Rotation Analysis of Variance (ANOVA) was conducted. The main effect for Rotation was significant, $F(8, 2818) = 33.49$, $p < .001$, indicating that the ratings differed by rotation. The main effect for Task was also significant, $F(9, 2818) = 18.88$, $p < .001$, indicating that there were differences in task ratings. The Rotation by Task interaction effect was significant, $F(72, 2818) = 1.62$, $p = .001$, indicating that the task ratings varied by task over rotations. For example, the lowest rated

task (M=1.8) for platoons in Rotation 2 was "Construct Obstacles", but it was the highest rated task for platoons in Rotation 9 (M=3.2).

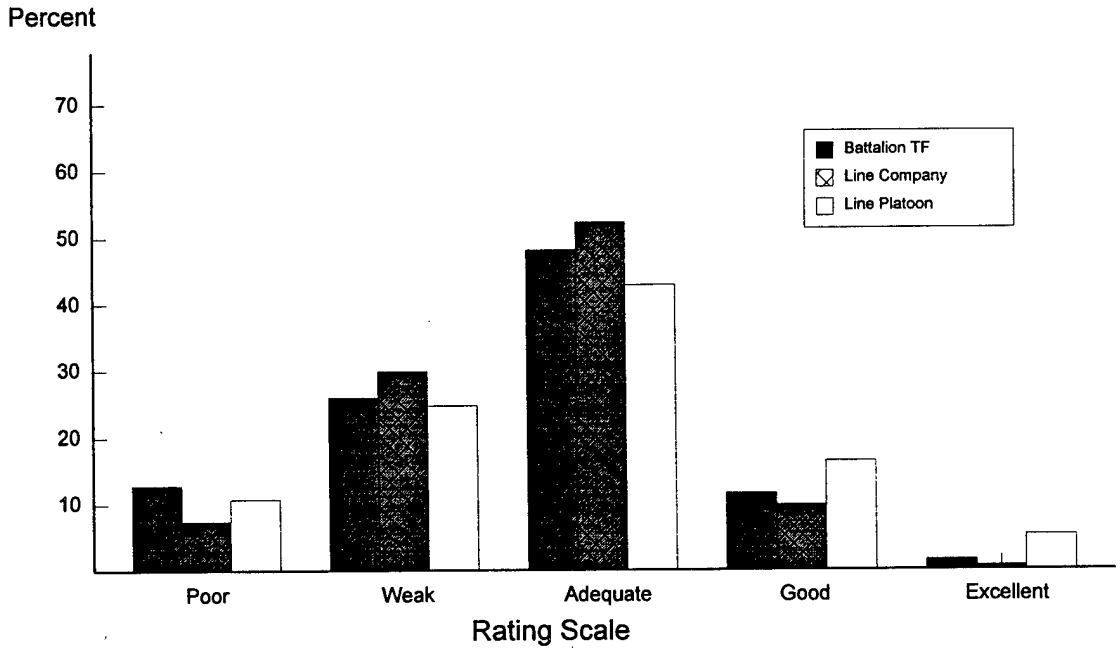

Figure 2. Percent ratings on subtasks by echelon.

Although the analysis of variance indicated that researchers can perform statistical analyses which will provide task discrimination, the practical significance must be questioned. The main effect for task was significant indicating that there were differences among the tasks. Examination of the "all" column which is a mean of each task for all rotations, reveals that the task ratings ranged from 2.3 to 3.3. On the five point scale this range would correspond to a little better than weak to a little better than adequate. Conclusions based on one point variance for ten tasks on a five point scale are tenuous at best. Means of the subtask scores for the same tasks were calculated to determine whether subtask score means vary more than the overall task means.[1] The variance of the subtask score means was similar to that of the overall task means ranging from a low of 2.5 for "Defend" to a high of 3.4 for "Helicopter Movement". No further analysis was conducted on the subtask scores.

[1] O/Cs rate subtask scores and the overall task score separately. The overall task score in the data base is not a mean of the subscores.

Table 3

Platoon Means for Ten Most Frequently Rated Tasks

	Rotations									
Platoon Tasks	1	2	3	4	5	6	7	8	9	all
Helicopter Movement	2.4	4.0	3.0	2.8	3.5	3.7	3.0	3.3	3.2	3.3
Defense	1.9	2.2	2.3	2.3	2.5	2.3	2.7	2.6	2.9	2.3
Occupy Area	2.3	2.9	2.5	2.3	2.2	2.2	2.6	2.9	2.6	2.5
Move Tactically	2.7	3.5	3.1	2.9	2.8	2.7	3.1	3.1	3.1	3.0
Cross Danger Area	2.7	3.5	2.7	2.5	2.4	2.4	2.7	3.1	2.8	2.9
Consolidate & Reorganize	2.8	3.4	3.3	2.9	2.4	2.4	2.9	2.9	2.8	2.9
Employ Fire Support	2.5	3.2	2.2	2.3	2.5	2.1	2.3	2.7	2.6	2.6
Construct Obstacles	2.7	1.8	2.6	2.4	2.7	2.2	2.7	2.9	3.2	2.6
Sustain Operations	2.8	3.6	2.8	2.8	2.9	2.6	2.8	3.0	2.9	3.0
Prepare for Combat	2.5	3.0	2.8	2.7	2.7	2.5	2.8	3.0	2.8	2.7

Company task means were calculated for the thirteen most frequently rated tasks (see Table 4). The procedure for calculating the company means was the same as the procedure for calculating platoon means. A Rotation by Task analysis of variance was conducted on the overall task means. The main effect for Rotation was significant, $F(8,1054) = 3.59$, $p < .001$, indicating that the task ratings differed over rotations. The main effect for Task was significant, $F(12, 1054) = 5.39$, $p < .001$, indicating that tasks ratings differed. The Rotation by Task interaction was not significant $F < 1$, indicating that the relationships among the tasks remained constant across rotations.

Again, a statistical analysis indicated that company tasks differ. This could lead some researchers to conclude that there is sufficient task discrimination to provide information on training focus. However, the practical significance must be questioned when an examination of the "all" column in Table 4 reveals that the overall task means varied from 2.3 to 2.9. This corresponds to all tasks being slightly better than weak to not quite adequate. The subtask score means were calculated for the same tasks with similar results. The lowest rated task was "Occupy an Assembly Area" (M=2.2) and the highest rated task was "Perform Personnel Actions" (M=2.9). No further analyses were performed on the subtask scores.

Based on the statistical examinations of platoon and company task means, it would appear that task discrimination analyses could be performed. Tasks in the upper 25% could be identified as strengths for typical units, whereas tasks in the lower 25% could be identified as weaknesses for typical units. However, because of the low variation in task means, it does not appear that this would make much sense from a practical viewpoint. That is, would anyone really believe that tasks rated slightly better than weak are in any more need of training than tasks that are rated slightly worse than adequate?

Table 4

Company Task Means for Thirteen Most Frequently Rated Tasks

	Rotations									
Company Tasks	1	2	3	4	5	6	7	8	9	all
Assault	2.6	2.5	2.3	2.5	2.0	2.5	2.3	2.4	3.1	2.5
Defend	2.3	2.0	2.2	2.3	2.7	1.8	2.3	2.2	2.6	2.3
Occupy an Assembly Area	2.7	2.3	2.4	2.6	2.7	2.0	2.2	2.5	2.0	2.4
Move Tactically	2.5	2.6	2.6	2.5	3.0	2.7	2.4	2.8	2.6	2.7
Consolidate & Reorganize	2.6	2.9	2.7	2.0	2.6	2.3	2.0	2.5	2.3	2.5
Actions on Contact	2.4	2.8	2.6	2.5	2.6	2.3	2.3	2.8	2.6	2.6
Perform Logistics Sustainment	2.8	2.6	3.0	2.7	2.9	2.9	2.4	2.7	2.9	2.8
Perform Personnel Actions	3.0	2.5	2.8	2.9	3.2	2.9	2.7	2.8	2.9	2.9
Defend Against Air Attack	2.6	2.0	2.7	2.0	2.8	2.3	2.0	2.8	2.7	2.5
Employ Fire Support	2.9	3.1	2.5	3.0	2.8	2.4	2.6	2.8	3.0	2.8
Install, Operate, & Maintain Radio	2.7	2.9	3.1	2.6	3.0	2.9	2.5	2.8	3.1	2.9
Develop & Communicate a Plan	2.6	2.7	2.4	2.8	2.8	2.3	2.2	2.5	2.8	2.6
Maintain OPSEC	2.7	2.7	2.7	2.6	2.9	2.5	2.4	2.5	2.8	2.6

Subtask discrimination. Another possible use of the T&EO data base is to examine subtask scores on selected tasks to determine what factors contribute to the success or failure for those tasks. A past argument for rating the subtasks has been that the added detail will allow researchers this opportunity. The subsequent benefit would be a training focus for selected tasks. For the purpose of this study, the tasks "Linkup" and "Consolidate and Reorganize" were selected from both the platoon and company data. These tasks were selected simply because they were frequently rated tasks, allowing sufficient data for analysis. Additionally, the tasks are performed at both platoon and company levels. Subtask means were calculated for all four tasks using all the ratings for all nine rotations. One-way ANOVAs were conducted for all four tasks on the subtask scores.

The ANOVA for the platoon task "linkup" was not significant, $F(3,469) = 2.19$, $p > .05$. Table 5 contains the means and the verbal descriptions of the four subtasks for "linkup". The means varied from a low of 2.6 to a high of 2.9. Not only was there no statistically significant difference, but it appears that there would be little possibility of the subtask means leading to any useful conclusions regarding task performance.

The ANOVA for the platoon task "Consolidate and Reorganize" was significant, $F(15, 4099) = 13.92$, $p < .001$. Table 6 contains the means and verbal descriptions of the 16 subtasks associated with the task. Although the subtask scores were statistically significantly

different, the means ranged from a low of 2.6 to a high of 3.4. Again, it does not appear that there would be any useful information contained within the subtask scores which would provide a training focus.

Table 5

Subtask Means for the Platoon Task "Linkup".

Subtask verbal description	Mean
The platoon leader identifies the tentative linkup site by map reconnaissance or a linkup site is designated by higher headquarters.	2.9
The platoon leader coordinates or obtains information from the unit that his platoon will link up with.	2.6
The stationary unit performs linkup actions.	2.6
The moving unit performs linkup actions.	2.6

Table 6

Subtask Means for the Platoon Task "Consolidate and Reorganize".

Subtask verbal description	Mean
Platoon leader positions or repositions OP forward to provide security.	2.6
Platoon occupies or reoccupies hasty fighting positions near the objective & establishes security.	2.8
Leaders adjust positions and position crew-served weapons.	2.9
Platoon searches area to ensure it is free of the enemy.	2.7
Platoon leader assigns or reassigns all squads temporary sectors of fire.	2.8
Squad and team leaders adjust positions to cover likely avenues of approach.	2.9
Platoon prepares, replaces, or repairs fighting positions and obstacles, as needed.	2.6
The platoon re-establishes the chain of command & communications net.	3.2
The platoon fills key positions in the following priority:	3.2
Leaders supervise distribution of ammunition and equipment.	3.0
Squad leaders report ammo, personnel, PWs, & equipment status to Platoon leader & request medical help.	3.1
Platoon leader reports status of & requests replacement for personnel, weapons, ammo, & equipment.	3.0
Platoon leader collects & disseminates information about the completed operation.	2.9
All PWs are handles IAW the 5 Ss and are tagged.	2.9
Casualties are treated.	3.4
Casualties are evacuated.	3.1

The ANOVA for the company task "linkup" was not significant, $F(4, 52) < 1$. The subtask means associated with the task are shown in Table 7. The subtask means varied from a low of 2.1 to a high of 2.3. As with the two platoon tasks, it does not appear that the subtask scores provide any additional information which would be useful in determining training focus.

Table 7

Subtask Means for the Company Task "Linkup".

Subtask verbal description	Mean
The commander identifies tentative, primary, & alternate linkup sites by map reconnaissance or a linkup sites are designated by higher headquarters.	2.3
Based on estimate and METT-T, commander develops a linkup plan.	2.2
Commander coordinates to obtain information from the linkup unit.	2.1
Stationary unit actions	2.2
Moving unit actions.	2.2

The ANOVA for the company task "Consolidate and Reorganize" was not significant, $F(9, 567) = 1.46$, $p > .05$. Table 8 contains the means for the ten subtasks associated with the task. The means ranged from a low of 2.1 to a high of 2.5. Consistent with the other three tasks, the additional information gained from the subtask scores would be minimal to researchers.

Table 8

Subtask Means for the Company Task "Consolidate and Reorganize".

Subtask verbal description	Mean
Commander positions and repositions platoons to provide security to ensure main body is not engaged without warning.	2.5
Company eliminates enemy resistance and destroys, captures, or forces withdrawal of enemy with the company AO.	2.5
Maintain pressure on withdrawing enemy with direct/indirect fires on last known/ suspected positions.	2.4
Platoons place security elements forward of perimeter to provide early warning for main body.	2.1
Support elements are moved forward & integrated into the defense.	2.5
Commander adjusts/directs adjustment of positions and units.	2.4
Company searches the area to ensure it is free of enemy elements.	2.2
The commander assigns platoon sectors.	2.4
Conduct consolidation and reorganization activities.	2.5
The company is prepared to continue the mission.	2.4

Overall, there was slight subtask variation within the four selected tasks. Therefore, researchers may not be able to determine strengths and weaknesses associated with particular tasks. At the overall task level, tasks varied somewhat and differed statistically. Examination of the subtasks in an attempt to get at the causal effects or to focus training for particular tasks is not feasible given the little variation in the subtask ratings. In other words, it does not appear that any inferences can be drawn on why a particular task is weak or strong.

O/C biases. Another possible use of the T&EO data base would be for researchers to make unit comparisons based on variables like training innovations or doctrinal changes. To do this, the data must be summarized by rotation. Data prior to a training inovation or doctrinal change could be used as baseline data to compare to units in rotations after a change in training strategy or doctrine. To test whether the present data could be used for making unit comparisons or in providing unit feedback, the data were summarized by rotation. A major problem with using the data base in this way was pointed out by Fober (1993). The old data base contained apparent biases in the ratings by some O/Cs . Bias should be thought of as the overall tendency of an O/C to rate performance consistently high or low, or the tendency to limit ratings to a small range (i.e., failure to use the entire scale). For example, one specific platoon (e.g., A Company, 1st Platoon) was rated the highest platoon over several rotations, indicating that the O/C for that platoon was an "easy" rater. Fober (1993) came to this conclusion because O/Cs are assigned to rate the same platoon (e.g., A company, 1st platoon) during each rotation. If the ratings remain similar and by their pattern, different from other O/C's ratings over several rotations, an inference can be made that the O/C's ratings may have been artificially raised or lowered. These conclusions are somewhat tentative because it can not be determined with certainty when an O/C change takes place. However, it is reasonable to make assumptions when the same element is a consistent outlier over several rotations. The platoon means were calculated for platoons by rotation. Because there were two task forces, there were 18 line platoons. The platoon means are presented in Table 9.

Examination of Table 9 reveals that even with the new rating scale, there is the possibility for biased ratings. The shaded area of Table 9 highlights ratings from the same platoon (e.g., B company, 2nd platoon) over five rotations. The rating means appear to flagrantly violate common sense. Not only were the ratings unusually high (M = 4.4 to 4.9 on a five-point scale), but they were high for five consecutive rotations indicating likely biases on the part of the platoon O/C. Many O/Cs feel that the T&EO data go to a "black hole" (personal communications with several former O/Cs) and are never used. Therefore, many O/Cs view the T&EO ratings as unimportant. The O/C from Platoon 1 may have been testing the system or he may have truly been a biased rater. If he never received feedback for unusually high ratings, then his data remain suspect. Examination of the rest of the data does not reveal any apparent biases, however it is possible that many rating biases exist which are not as easily detectable as those from Platoon 1. No statistical analyses can be performed after the fact because there is no way to determine when O/Cs change out. However, standard deviations for each platoon may provide some indications of O/C biases. The assumption is that the "tighter" the scores, the more restrictive the O/C is in using the scale. Again, there is no way to know for sure whether the O/C is biased, but low standard deviations could be cause for concern. The standard deviations were

Table 9

Platoon Means of O/C Ratings Over Nine Rotations for Two Task Forces

	Rotation Number								
Platoon*	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
1	3.0	3.7	2.9	3.4	4.4	4.4	4.9	4.9	4.9
2	3.3	3.9	1.9	2.6	2.8	2.8	2.7	2.9	2.8
3	3.1	3.0	2.9	3.0	2.6	2.1	2.9	3.3	3.0
4	2.8	2.9	2.9	3.0	2.8	2.3	2.5	2.5	2.7
5	2.4	2.4	2.6	3.3	3.6	2.8	3.1	3.4	2.4
6	2.5	3.4	2.4	2.7	2.9	2.7	2.9	2.7	2.8
7	2.9	3.0	3.1	2.6	2.7	2.8	2.4	3.0	3.3
8	2.0	4.3	2.2	2.1	2.4	2.5	2.7	2.9	2.9
9	2.7	2.9	2.8	3.4	2.5	2.1	2.3	3.0	2.3
10	3.0	4.0	3.2	3.5	3.4	2.9	3.0	3.0	3.7
11	2.9	3.4	2.7	2.7	2.3	2.7	2.8	3.5	2.9
12	1.8	3.3	2.0	2.2	2.6	2.7	2.4	2.3	2.4
13	2.1	2.4	2.7	2.8	3.1	2.9	2.9	3.0	2.8
14	2.9	3.4	1.8	3.2	2.6	2.0	3.0	2.7	2.8
15	2.3	2.6	3.1	2.1	2.3	2.2	2.1	2.5	2.1
16	2.8	4.3	2.9	1.6	2.3	1.2	3.0	2.8	2.6
17	1.7	1.9	3.1	1.9	2.8	1.8	2.4	1.9	2.4
18	3.0	3.7	2.5	2.4	2.7	2.8	2.0	2.6	2.6

*Note. The platoon numbers were randomly assigned to assure confidentiality. However, the platoon numbers remain constant across rotations.

calculated along with the means shown in Table 9. For demonstration purposes, standard deviations of less than .60 are discussed in terms of possible cause for concern regarding possible O/C biases. The standard deviation of .60 was chosen arbitrarily, however, it was thought that standard deviations as low as only a half point on a five-point scale might be cause for concern. The standard deviations for each platoon over all rotations are presented in Table 10.

The shaded area of Table 10 highlights the same rotations and platoon which were identified as possible cause for concern based on the mean scores presented in Table 9. Based on the standard deviations seen in Table 10 along with the means seen in Table 9, it appears that the final three rotations are very likely a product of some O/C bias. Not only are the means high, but the standard deviations are low, indicating that the O/C was possibly restricting his use of the scale. Bias is a likely explanation given the ratings of the other platoons. It is not probable that the O/C assigned to Platoon 1 observed performance at such a high caliber for three straight rotations. The results can not be attributed to a lack of numbers available. There were 250-300

Table 10

Platoon Standard Deviations of O/C Ratings Over Nine Rotations for Two Task Forces.

	Rotation Number								
Platoon	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
1	.35	.54	1.55	1.83	1.18	1.16	.58	.34	.45
2	.81	.51	.93	.76	.58	.57	.50	.33	.48
3	.32	.26	.80	.72	.71	.96	.49	.65	.62
4	.58	.56	1.04	.78	.65	.84	1.17	1.01	.82
5	.75	.82	.81	1.02	.54	.85	.74	.84	.83
6	1.08	.89	.99	.83	.63	.59	.51	1.14	1.03
7	.87	.76	.64	.65	.58	.87	.66	.65	.70
8	.91	.67	.56	.80	.57	.72	.61	.55	.40
9	.44	.95	1.31	1.11	1.23	.61	1.03	.95	.46
10	1.38	.52	.92	.82	.69	1.17	.80	.98	.65
11	.75	.86	1.01	.74	.99	.61	.85	.77	.35
12	.77	.69	.88	.90	.83	.98	.75	.90	.85
13	.74	.51	.87	.72	.72	.62	.59	.66	.73
14	.82	.73	.77	.75	1.09	.99	.81	.67	.77
15	.70	.66	.86	.80	.69	.67	.92	.84	.82
16	.78	1.07	.73	.72	.61	.53	.65	.44	.69
17	.60	.71	1.21	.70	.80	.69	.74	.89	.92
18	.62	.64	.95	.84	1.03	.69	.86	.70	.52

subtask ratings for Platoon 1 during rotations 7, 8, and 9. Overall, there were 32 times where the standard deviation was lower than .60. These results are an indication that the O/Cs may not have been using the rating scale in the way in which it was intended. For example, Platoon 3 in Rotation 1 shows a mean of 3.1. The standard deviation is .32. In addition, the minimum was 2 with a maximum of 4 for 438 subtask ratings, indicating that there were a lot of 3 ratings.

Possible Solution

The results presented in the previous section indicate that there may be rating biases by some O/Cs which may impact on the task performance ratings. However, that does not necessarily mean that the data are useless to the researcher. For example, research examining overall task performance (e.g., trendlines), may not be affected by possible rating biases. As long as O/Cs are consistent in their ratings, trendline type analyses should not be greatly affected. A particular O/C may rate similar performances higher than another O/C, but the relationship among tasks is consistent. In other words, a poorly performed task will be one of the lowest rated tasks, regardless of whether the O/C is a "high" rater or not. However, this argument does not hold true; results presented earlier indicate that it is difficult to discriminate among the task

ratings. Therefore, the individual's tendency to restrict his rating scale is probably impacting on most research applications associated with the T&EO data base.

One of the possible contributing problems to the quality of the data base was the high number of ratings required of the O/Cs (Fober, 1993). Because their workload is already high and the priority on the T&EO ratings has been low, O/Cs will naturally complete the ratings as quickly as possible. For example, one platoon mission may have twenty possible tasks to rate. Although the O/C workload was reduced by dropping the requirement to rate subtask standards, many of the tasks may still have 20-30 subtasks to rate. A more manageable level of detail may need to be around five or six ratings for each task.

One platoon task, "Prepare for Combat", was chosen as an example of how the number of ratings might be reduced without reducing the level of detail needed by researchers. "Prepare for Combat" contains twenty-three subtasks, many of which should be related to each other. For home station training, it is important for a unit to complete training to standard on all subtasks. However, a performance measurement system may require more general categories. The reduced workload may increase the reliability of the ratings, allowing researchers to glean information about tasks from several general categories. These general indicators would then drive possible directed studies attempting to provide further detail.

To demonstrate how a task might be modified for use in performance measurement, a Subject Matter Expert (SME) from the U.S. Army Infantry School/Center (USAIS/C) at Fort Benning determined six categories which comprise the task "Prepare for Combat". The six categories chosen to possibly represent the task were:

- Supervision
- Information Updates
- Information Collection
- Analysis
- Direction (giving and receiving)
- Coordination

Seventeen SMEs from the USAIS/C were then tasked to examine the categories (verbal descriptions) and place each subtask from "Prepare for Combat" into the category which best described it. The SMEs were instructed to use only the verbal categories provided and to make each subtask fit into one category. Two subtasks were deleted from the analysis because there are normally insufficient data (e.g., "Vehicles are combat loaded IAW ...." and "All personnel will test-fire weapons if situation permits"). The remaining twenty-one subtasks were provided to the SMEs in the order in which they occur in the MTP. The SMEs were all small group instructors from the Infantry Officer's Advanced Course. Many of them had served as O/Cs at a Combat Training Center. The ratings of the SMEs were then examined to determine how much the SMEs agreed with each other on placing the subtasks in the general categories. Table 11 shows where the SMEs placed each subtask(Refer to Table 2 for the full verbal descriptions of each of the subtasks associated with "Prepare for Combat").

Examination of Table 11 reveals that in general the SMEs grouped the subtasks into the same categories. All of the twenty-one subtasks represented were placed in a category which

Table 11

SME groupings by subtask

Subtask Description	Supervision	Information Updates	Analysis	Collect Info	Direction	Coord-ination	Percent Agreement
Receive Mission	0	1	1	2	13	0	76%
Mission Analysis	0	0	0	17	0	0	100%
Complete Analysis	0	2	14	1	0	0	82%
Warning Order	1	2	0	0	13	0	76%
Readiness	15	1	0	0	1	0	88%
Tentative Plan	1	0	15	0	1	0	88%
Movement	1	0	0	1	10	5	59%
Recon	0	2	0	15	0	0	88%
Complete Plan	1	3	12	0	1	0	71%
OPORD	0	1	0	0	15	1	88%
Coordinate	0	0	0	0	0	17	100%
Mission Prep	15	0	0	0	0	2	88%
Sustain	0	0	10	0	2	5	59%
Continuous Recon	0	2	1	14	0	0	82%
Monitors	0	5	0	2	0	10	59%
Modifies Plan	1	4	2	0	10	0	59%
FRAGOs	1	2	0	0	14	0	82%
React to Orders	0	1	0	0	15	1	88%
Coordinate Actions	0	0	0	0	0	17	100%
Report Info	0	12	0	0	0	5	71%
Distribute Info	1	15	0	0	1	0	88%

was agreed on by the majority of the SMEs. This is not to say that these are the best categories to use, however, this quick demonstration occurred with minimal effort and time involved. Therefore, it might be used as a way to reduce the number of ratings required for tasks, thereby reducing the overall O/C workload.

These results suggest that the number of items to rate in a performance assessment system could be reduced by creating general categories. SMEs can make the choices as to where the subtasks fit within the categories. In the sample task, the ratings indicated that the O/Cs agreed on the general category descriptions for all the subtasks. The implications are that for the task "Prepare for Combat" the performance checklist could contain four to six categories instead of the current twenty-three. It should be stressed that this suggestion is only for performance assessment and not for training. Fewer items to rate would decrease O/C workload which in turn might increase the quality of the rating. That is, O/Cs might be more apt to critically assess a few items as opposed to "checking the block" when there are too many items to rate.

General Discussion

The original purpose of the T&EO data base was to provide researchers with detailed information on unit performance to complement the unit Take Home Packages. Whereas Take Home Packages provide a narrative of unit performance sufficient for unit feedback, the T&EO data base should be easily accessible and easily manipulated (i.e., data base is coded) so that researchers can not only describe unit performance, but be able to make inferences as well. The T&EO data base should also be of sufficient detail to provide information which can not be found within the format of the Take Home Packages.

Data Base Assessment

Examination of the overall ratings summarized across all echelons indicated that for this sample, the entire rating scale was used. Further analysis at the task level revealed that there were statistical differences in task performance both between tasks and within tasks over rotations. However, from a practical standpoint researchers would have difficulty making recommendations and conclusions based on differences less than one point. The usefulness of the data is further reduced when the possibilities for O/C biases are taken into account. Since O/Cs fill out the T&EOs after the rotation, they may restrict their ratings on the subtasks based on their overall impression of the unit. The demonstration using SMEs to reduce the number of ratings was encouraging in that for most tasks the subtasks could be grouped into more general categories (e.g., coordination). This procedure could be used as a model for further reducing the number of rated items within tasks, thereby possibly increasing the quality of the ratings.

Conclusion and Recommendations

The changed format provided data which on face value were of increased usefulness over the prior system. Task discrimination was possible using statistical analyses both within and between tasks. Again, the usefulness of the statistical results is probably very limited. There are

still possible problems resulting from O/C biases and possible lack of command emphasis on the T&EO ratings. In part because of the problems associated with the T&EO data base, JRTC discontinued collecting these data after the 95-7 rotation.

The T&EO data base had potential to provide researchers with the required data to conduct meaningful studies. If a performance measurement system is ever again adopted at JRTC, the T&EO data base could be used as a model with some necessary changes. The number of items (subtasks) for each task must be reduced. A panel of SMEs could accomplish this in advance in a process similar to that done in the present study for "Prepare for Combat". In addition, the rated items should enable the O/Cs to use the data for their own use (e.g., After Action Reviews and Take Home Packages). To make this happen, current job aids and "cheat sheets" used by O/Cs must be examined along with O/Cs' input. This information could be synthesized for use as the performance measurement system. Command emphasis and periodic feedback must be provided to the O/Cs to ensure that the data reflect unit performance. The feedback provided to the O/Cs must not only identify possible rating problems, but it must also show that the data are being used for meaningful purposes. If the data are just going to a "black hole", then the decision to discontinue collecting T&EO data was a wise one. Researchers will have to develop other means or use the existing narrative data (e.g., THPs) as measures of unit performance.

References

Department of the Army. (1988). Mission training plan for the Infantry rifle platoon and squad (ARTEP 7-8-MTP). Washington, DC: Author.

Fober, G. W. (1993). The Joint Readiness Training Center's training and evaluation outline data base: Preliminary assessment (ARI Research Note 93-11). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A260 128)

Fober, G. W., Dyer, J. L., & Salter, M. S. (1994). Measurement of performance at the Joint Readiness Training Center: Tools of Assessment. In R. F. Holz, J. H. Hiller, & H. H. McFann (Eds.), Determinants of effective unit performance: Research on measuring and managing unit training readiness, 39-67. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Nichols, J. J. (1991). Analysis of the scope, functionality, and usability of the Joint Readiness Training Center (JRTC) data base archive (Research Note 91-24). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A232 128)